International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

AI -Powered Web Application for Automated Short Video Generation

R. Naveen Kumar
UG Scholar,
Vels Institute of Science,
Technology And Advanced Studies (VISTAS),
Pallavaram, Chennai-600117,
Tamil Nadu, India.
rnkk9912@gmail.com

Dr. AS. Arunachalam, MCA., M.Phil.,Ph.D.,
Professor,
Vels Institute of Science,
Technology And Advanced Studies (VISTAS),
Pallavaram, Chennai-600117,
Tamil Nadu, India.
arunachalam1976@gmail.com

https://doi.org/10.55041/ijsmt.v2i4.618

Cite this Article: Kumar, R. N. (2026). AI -Powered Web Application for Automated Short Video Generation. International Journal of Science, Strategic Management and Technology, *02*(05). https://doi.org/10.55041/ijsmt.v2i4.618

ABSTRACT: In the rapidly evolving digital landscape, video content has become a dominant medium for communication, education, entertainment, and marketing due to its high engagement and effectiveness. However, traditional video production is a complex, time-consuming, and resource-intensive process that requires significant technical expertise. To address these challenges, this project presents the design and development of an AI-powered one-minute text-to-video generation web application that automates the creation of short, high-quality videos from user-provided textual inputs. The proposed system integrates advanced Artificial Intelligence techniques, including Natural Language Processing (NLP), Text-to-Speech (TTS), and diffusion-based generative models, to convert textual descriptions into dynamic video content. The input text is processed into semantic representations and structured into meaningful scenes, ensuring logical flow and coherence. For each scene, relevant visual content is generated using AI-based image and video synthesis, while voice narration is produced through TTS systems. Additional multimedia elements such as subtitles, transitions, and background music are incorporated to enhance the quality and effectiveness of the generated video. The system is built upon a pretrained video diffusion pipeline that iteratively refines latent representations to produce temporally consistent video frames. These frames are then encoded into standard video formats using multimedia processing techniques. The application is developed with a user-friendly web interface and a robust backend powered by modern deep learning frameworks.

To ensure efficient performance, optimization techniques such as mixed-precision computation, memory-efficient processing, and GPU acceleration are employed. These enhancements enable faster inference and improved scalability, making the system suitable for deployment in GPU-enabled environments. The proposed solution significantly reduces manual effort, production time, and cost, thereby making video creation accessible to students, educators, content creators, and businesses. This project demonstrates the practical application of generative artificial intelligence in multimedia content creation and highlights the transition from static image synthesis to automated video generation. Despite its advantages, challenges such as computational complexity , hardware

Keywords
Artificial Intelligence, Text-to-Video Generation, Natural Language Processing, Diffusion Models, Text-to-Speech, Video Synthesis, Web Application, Multimedia Processing, Generative AI, Automation

I . INTRODUCTION

The rapid advancement of Artificial Intelligence (AI) has revolutionized the process of digital content creation, particularly in the domains of image synthesis, audio generation, and video production. Among these emerging technologies, text-to-video generation stands out as a transformative innovation that enables machines to create dynamic visual content directly from natural language descriptions. This paradigm shift reduces the

dependency on traditional content creation methods and introduces a new era of automated multimedia generation.

This papert focuses on the design and development of a Text-to-Video Generation Web Application, which allows users to generate short, meaningful video clips by simply providing descriptive text prompts. The system aims to bridge the gap between human imagination and digital visualization by leveraging state-of-the-art generative AI models. Unlike conventional video production workflows—which involve multiple stages such as scripting, shooting, editing, and rendering—this approach automates the entire pipeline, significantly reducing the time, cost, and technical expertise required.

Concept of Text-to-Video Generation Using AI : Text-to-video generation is an advanced application of generative AI that combines Natural Language Processing (NLP) with computer vision techniques to convert textual input into a sequence of visual frames. At the core of this technology are diffusion models, which have recently gained popularity due to their ability to generate high-quality and realistic outputs. These models operate by starting with random noise and iteratively refining it into structured and meaningful data through a denoising process. In this project, a diffusion-based architecture is employed to generate temporally consistent video frames. The input **text** prompt is first transformed into semantic embeddings, capturing the contextual meaning of **the** generate short video clips with coherent motion and structure.

Another significant contribution in this domain is the integration of transformer-based text encoders, such as CLIP, which convert textual prompts into embeddings that guide the generation process. These embeddings ensure that the generated video aligns closely with the user's input description.

Recent advancements, including large-scale models like CogVideoX, have further improved video quality and realism but require substantial computational resources. As a result, lightweight alternatives like ZeroScope are often preferred for practical implementations.

In addition to model development, several studies have focused on optimizing inference performance using techniques such as mixed-precision computation, memory-efficient attention mechanisms, and model parallelism. These optimisations are crucial for deploying such systems in real-world environments.

Overall, the literature indicates a clear evolution from GAN-based approaches to diffusion-based architectures, with current research focusing on improving efficiency, scalability, and video realism. This project builds upon these advancements by implementing a practical text-to-video system using a diffusion-based model within an accessible web interface.

II. LITERATURE SURVEY

The field of generative AI has seen rapid advancements over the past decade, particularly in the domain of image and video synthesis. Early approaches to visual content generation primarily relied on Generative Adversarial Networks (GANs), which consist of a generator and a discriminator working in opposition. While GANs demonstrated promising results in image generation, they often suffered from issues such as training instability, mode collapse, and difficulty in generating temporally consistent video sequences.

To overcome these limitations, researchers introduced diffusion models, which have recently gained popularity due to their stability and superior output quality. Diffusion models operate by gradually adding noise to data and then learning to reverse this process. Notable works such as Denoising Diffusion Probabilistic Models (DDPM) have laid the foundation for modern generative systems. These models have been successfully applied in text-to-image tasks, as seen in systems like Stable Diffusion.

Extending diffusion models to video generation introduced additional challenges, particularly in maintaining temporal consistency across frames. To address this, researchers developed Video Latent Diffusion Models (VLDMs), which incorporate temporal attention mechanisms into the architecture. Models such as ZeroScope and ModelScope-based video diffusion systems demonstrate the ability to

pretrained diffusion model, specifically the ZeroScope V2 576w model. This model operates in a latent space and generates video frames by iteratively refining random noise into meaningful visual representations. The input text is encoded into embeddings, which guide the generation process to ensure that the output aligns with the provided description. The model also maintains temporal consistency across frames, resulting in smooth and realistic video sequences.

Video Rendering and Output

After the generation of frames, the system compiles them into a standard video format using multimedia processing techniques. This stage includes frame sequencing, encoding, and optional enhancements such

as audio integration and subtitles. The final video output is then delivered to the user via the frontend interface.

System Workflow

The system follows a sequential workflow: the user inputs a text prompt, the backend processes and encodes the input, the AI model generates video frames using a diffusion process, and the frames are compiled into a video that is displayed to the user. This streamlined process ensures efficient and automated video generation.

III SYSTEM ARCHITECTURE

The proposed AI-powered Text-to-Video Generation Web Application is designed using a modular and layered architecture to efficiently convert user-provided textual input into a coherent video output. The system integrates web technologies with advanced deep learning models to ensure scalability, usability, and performance. The architecture primarily consists of three major components: the user interface layer, the backend processing layer, and the AI-based video generation layer.The overall workflow begins with the user providing a text prompt through the web interface. This input is forwarded to the backend server, where it undergoes preprocessing and semantic encoding. The processed input is then passed to the diffusion-based AI model, which generates a sequence of video frames. These frames are compiled into a video format and returned to the user through the interface.

User Interface Layer : The user interface layer is developed using Gradio, which provides an interactive and intuitive platform for user interaction. This layer enables users to input textual prompts and views the generated video output in real time. It acts as a bridge between the user and the system, ensuring ease of access and usability without requiring technical expertise.

Backend Processing Layer : The backend processing layer is implemented using Python and is responsible for handling the core logic of the application. It performs input validation, preprocessing, and conversion of text into semantic representations. The backend communicates with the AI model using deep learning frameworks such as PyTorch and utilizes the Hugging Face Diffusers library to load and execute pretrained models. Additionally, optimization techniques such as mixed-precision computation and efficient memory management are applied to improve performance.

AI-Based Video Generation Layer : The AI-based video generation layer forms the core of the system and utilizes a

features such as background music, subtitles, and transitions can also be integrated at this stage to enhance the final output.

F. Output Module: The Output Module delivers the generated video to the user through the web interface. It displays the final video and allows the user to view or download it. This module ensures smooth playback and provides a satisfactory user experience.

G. System Optimization Module: This module is responsible for improving system performance and efficiency. It includes techniques such as mixed-precision computation, memory optimization, and GPU acceleration. These optimizations help reduce processing time and enable the system to handle computationally intensive tasks more effectively.

These embeddings guide the generation process, ensuring that the produced frames align with the user's description. The frames are then combined sequentially to form a coherent video clip. The system utilizes the Zero Scope V2 576w model, a latent diffusion-based model specifically designed for video generation tasks. This model is capable of maintaining temporal consistency across frames, which is a critical requirement for generating smooth and realistic videos. Additionally, the integration of Text-to-Speech (TTS) and other multimedia enhancements can further enrich the generated content, making it more engaging and expressive

System Implementation and Technological Integration: The implementation of the proposed system involves the integration of modern machine learning frameworks and web technologies to ensure both functionality and usability. The backend of the application is developed using Python and powered by PyTorch, which provides efficient support for deep learning model execution. The Hugging Face Diffusers library is utilized to access and manage pre-trained diffusion models, enabling faster development and deployment.

For the user interface, Gradio is employed to create an interactive and user-friendly web platform. This allows users to easily input text prompts, initiate video generation, and view the results in real time without requiring technical knowledge. The system architecture is designed to handle the computational demands of

video generation by incorporating optimization techniques such as mixed-precision computation and GPU acceleration.

Despite its advantages, the system faces certain limitations, including high computational requirements, dependency on hardware resources, and challenges in scaling for large numbers of users. These limitations highlight the need for further research in model optimization, efficient resource utilization, and scalable deployment strategies.

IV MODULES DESCRIPTION

The proposed AI-powered Text-to-Video Generation Web Application is divided into several functional modules to ensure a structured and efficient workflow. Each module is responsible for a specific task, enabling smooth data processing and system execution.

A. Input Module: The Input Module is the initial stage of the system where the user provides a text prompt or script through the web interface. This module is developed using Gradio, which offers an interactive and user-friendly environment. It captures the user's input and sends it to the backend for further processing. The module ensures proper validation of input data to avoid errors during execution.

B. Text Processing and Encoding: Module This module processes the input text and converts it into a machine-understandable format. Natural Language Processing (NLP) techniques are used to analyze the text and extract semantic meaning. The processed text is then transformed into embeddings, which serve as the guiding input for the AI model. This step ensures that the generated video content aligns with the user's description.

C. AI Video Generation Module: The AI Video Generation Module is the core component of the system. It utilizes a pretrained diffusion model, specifically the Zero Scope V2 576w model, to generate video frames. The module begins with random noise in a latent space and iteratively refines it into meaningful visual representations based on the encoded text input. This process ensures temporal consistency across frames, resulting in smooth and realistic video sequences.

D. Frame Processing Module: Once the frames are generated by the AI model, this module handles the organisation and processing of individual frames. It ensures correct sequencing, resolution consistency, and frame alignment. This module plays a crucial role in maintaining the visual quality and continuity of the video.

E. Video Rendering Module: The Video Rendering Module converts the sequence of processed frames into a standard video format such as MP4. It uses multimedia processing libraries to encode frames into a playable video. Additional which generates a sequence of video frames by refining random noise into meaningful visuals.

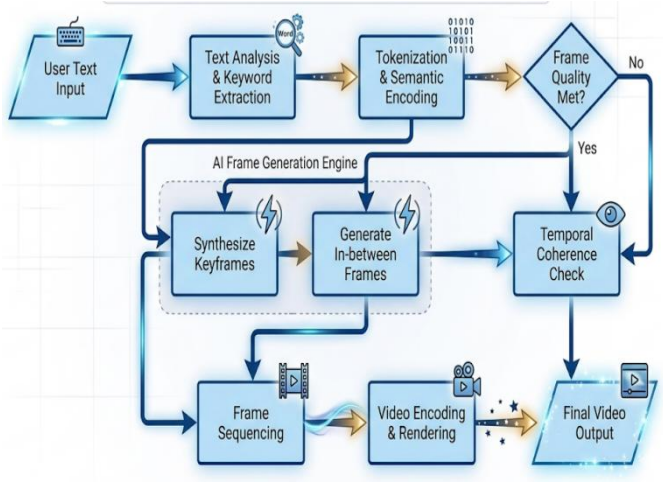
F. Frame_Sequencing: The generated frames are organized in a proper sequence to maintain continuity and smooth transitions.

G. Video_Compilation: The frames are combined and encoded into a video format such as MP4 using multimedia processing libraries.

H. Enhancement: Additional features such as voice narration, subtitles, and background music may be added.

I. Output_Display: The final video is displayed to the user through the interface, with options for playback and download.

Workflow Architecture



Tools Used:

The development of the proposed AI-powered Text-to-Video Generation Web Application involves the integration of various software tools, frameworks, and technologies to ensure efficient performance and scalability.

A. Programming Language

- Python is used as the primary programming language due to its simplicity, flexibility, and strong support for machine learning and web-based applications.

B. Deep Learning Framework

- PyTorch is utilized for implementing and executing deep learning models. It provides dynamic computation

graphs and efficient GPU acceleration for handling complex AI tasks.

C. AI Model and Libraries

- Hugging Face Diffusers library is used to access pretrained diffusion models for video generation.
- ZeroScope V2 576w Model is employed as the core model for generating temporally consistent video frames from textual input.

D. Natural Language Processing

- NLP techniques are used to process and encode input text into semantic embeddings, enabling the AI model to understand user prompts effectively.

E. User Interface

Gradio is used to develop the frontend interface, allowing users to input text prompts and visualize generated videos in an interactive manner.

dependency, and scalability remain areas for future improvement.

V METHODOLOGY AND TOOLS USED

Methodology : The methodology of the system is based on a pipeline approach, where the output of one stage becomes the input for the next stage. The entire process is divided into multiple phases to ensure clarity, efficiency, and modular execution. Initially, the system accepts a textual prompt from the user, which describes the desired video content. This input is then processed using NLP techniques to extract semantic meaning and context. The processed text is converted into embeddings, which act as a guiding representation for the AI model. The core of the methodology lies in the use of a diffusion-based video generation model. The model operates by initializing random noise in a latent space and gradually refining it into structured visual frames through multiple iterations. The semantic embeddings derived from the input text guide this refinement process, ensuring that the generated frames are aligned with the user's description. Once the frames are generated, they are processed and arranged sequentially to maintain temporal consistency. These frames are then encoded into a standard video format using multimedia processing techniques. Optional enhancements such as audio narration, subtitles, and transitions can also be added to improve the overall quality of the video. To ensure efficient execution, the system incorporates optimization strategies such as mixed-precision computation and GPU acceleration, which significantly reduce processing time and resource consumption.

Working Process: The working process of the system can be described step-by-step as follows:

1. User_Input:

The user enters a text prompt or script through the web interface.

2. Input_Processing:

The system validates and preprocesses the input text to remove errors and improve quality.

3. Text_Encoding:

The processed text is converted into semantic embeddings using NLP techniques.

4. Frame_Generation:

The encoded input is passed to the diffusion model**,**

F. Multimedia Processing

- Libraries for video processing are used to convert generated frames into standard video formats such as MP4, ensuring compatibility and smooth playback.

G. Hardware and Acceleration

- GPU acceleration (CUDA-enabled devices) is used to improve performance and reduce processing time.
- Mixed-precision computation is implemented to optimize memory usage and speed.

VI CONCLUSION

In this project, an AI-powered Text-to-Video Generation Web Application has been successfully designed and implemented to automate the process of video creation from textual input. The system integrates advanced technologies such as Natural Language Processing, diffusion-based generative models, and multimedia processing to produce short, coherent video clips.

The proposed solution significantly reduces the time, effort, and technical expertise required for traditional video production. By providing a user-friendly interface and efficient backend processing, the system makes video generation accessible to a wide range of users, including students, educators, and content creators.

The project also highlights the growing importance of generative artificial intelligence in digital content creation and demonstrates the transition from static image synthesis to dynamic video generation. Despite challenges such as computational complexity and scalability, the system provides a strong foundation for future enhancements.

Future improvements may include optimizing model performance, enhancing video quality, reducing processing time, and integrating additional features such as real-time generation and advanced editing capabilities**.**

REFERENCES

[1]. R. Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models," Proc. IEEE CVPR, 2022.

[2]. Hugging Face, "Diffusers Library Documentation," 2024. [Online]. Available: https://huggingface.co/docs/diffusers

[3]. Cerspense, "ZeroScope V2 Model," Hugging Face, 2024.

[4]. A. Vaswani et al., "Attention Is All You Need," Proc. NIPS, 2017.

[5]. PyTorch, "PyTorch Documentation," 2024. [Online]. Available: https://pytorch.org

[6]. Gradio, "Gradio: Build Machine Learning Web Apps," 2024. [Online]. Available: https://gradio.app

[7]. OpenAI, "ChatGPT: AI Language Model," 2025.

[8]. I. Goodfellow et al., "Deep Learning," MIT Press, 2016.

[9]. Arunachalam, A. S., and K. Rajeswari. "An Inclusive Survey of Student Performance With Various Data Mining Methods." *International Journal of Engineering and Technology (IJET) vol* 7 (2018): 522-525.